Supplementary Methods

Evaluating the impact of TAD variation on identification of collateral dependencies.

We undertook additional analyses to ensure that neither TAD size variability nor definitions would significantly impact the identification of fusion-associated collateral dependencies. To evaluate whether there is a correlation between the number of genes in a TAD and the odds of having a collateral dependency, we first stratified genes into rough quartiles of TAD size based on the number of genes in each TAD. There were 16,151 genes that were assigned to TADs based on our definitions. Edge cases of genes falling between quartiles were randomly assigned to one quartile or another (resulting in some variability in the number of genes assigned to each quartile). The final counts of genes attributed to each TAD quartile stratified by size were as follows:

TAD quartile	Gene count determining TAD sizes	Number of genes assigned to TAD quartile
1	**1 - 6**	**3,865**
2	**6 - 11**	**3,984**
3	**11 - 22**	**4,455**
4	**≥ 22**	**3,847**

We next used these quartiles to carry out a stratified enrichment analysis for collateral genes among dependencies in the context of fusions to demonstrate that the degree of enrichment of fusion collateral genes among dependencies was not significantly different between TAD quartiles.

We also sought to determine if using genomic distances, instead of TAD boundaries, would result in a different outcome for the enrichment of collateral genes amongst dependencies. Focusing on fusions exclusively, for each cell line, we 1) identified genes that were absolute dependencies and 2) identified genes that were “collateral” genes relative to fusions in that cell line, based on a symmetric genomic window of 930 kb (465 kb on either side of the start of a gene), corresponding to the average TAD size. Across all cell lines, we observed the total instances where an absolute dependency was a collateral gene vs. not a

collateral gene, calculated the odds ratio (OR), and determined significance using a Fisher's exact test to establish that the enrichment of collateral genes amongst dependencies in the context of fusions was robust to alternative TAD definitions.

Using fusions as a biomarker to identify associated overexpressed genes across DepMap through genome-scale screening. An analogous approach to that described for identifying fusion-associated dependencies was applied to identify genes overexpressed in a genome-scale pan-cancer unbiased screen. In order to enable uniform comparison between cell lines with and without a fusion of interest, we consistently evaluated RNA expression for all exons in protein-coding genes, as calculated by the Broad Institute's GTEX pipeline (https://github.com/broadinstitute/gtex-pipeline/blob/master/TOPMed_RNAseq_pipeline.md). Again, for each of the 3,277 fusions, all cell lines were stratified by the presence or absence of the fusion of interest, the mean $\log_2(TPM + 1)$ of RNA expression for each protein-coding gene was calculated for each group, and a two-sample t-test with the assumption of equal variance was carried out as a screen to identify genes that were overexpressed based on the difference in $\log_2(TPM + 1)$ between both groups. Correction for multiple-hypothesis testing was done using the Benjamini–Hochberg method to arrive at Q values. Partner and collateral genes were defined as above. We used thresholds of $\log_2$ fold change $(TPM + 1) > 1$ and $Q < .05$ to identify genes that were significantly overexpressed. Q values were again $-\log_{10}$ transformed for data visualization.

Further validation of enrichment of SV partner and collateral genes among absolute dependencies. To further validate findings from our enrichment analyses, we performed 209 iterations of the analyses, removing a single cell line at a time for each iteration to determine if any individual cell line was driving the observed enrichment. We also carried out multivariate logistic regression analyses to determine if the association between dependency status and partner status, as well as dependency status and collateral status, would persist when accounting for the covariates of disease type and cell line.

Hotspot driver mutation analysis in cell lines with and without fusion-associated dependencies. For the 645 cell lines for which fusion and dependency data was available, we evaluated each cell line for the presence or absence of a hotspot mutation in a gene from the COSMIC cancer census (https://cancer.sanger.ac.uk/cosmic/curation). To identify hotspot driver mutations, we narrowed our scope to known COSMIC Cancer Census genes with mutations that were recurrently seen in the TCGA (range 3 – 784 occurrences, mean 73 occurrences). Cell lines were subsequently stratified by the presence or absence of fusion-associated dependencies to compare the proportion of cell lines with hotspot driver mutations and the mean number of hotspot mutations per cell line.

Permutation testing to determine if fusion-associated differential dependencies occur more than would be expected by chance. To determine if fusion-associated differential dependencies were occurring more than would be expected by chance we took two different approaches to permutation testing: permuting gene labels and permuting fusion labels. For our gene-label permutation, for each of 3,277 fusions, we kept the total counts of significant dependencies, partners, and collateral genes constant. We shuffled gene labels relative to dependency probability scores 1,000 times for each fusion, **controlling for RNA expression** by only allowing for shuffling within a given RNA expression quartile, and counting the number of fusion-dependency pairings occurring by chance across all fusions with each iteration. For collateral fusion-dependency pairings, we carried out an additional gene-label permutation, this time **controlling for TAD size** by shuffling gene labels within TAD quartiles as defined before. For our fusion-label permutation, we kept each fusion-dependency relationship constant and shuffled fusion labels relative to fusion-dependency relationships 1,000 times, **controlling for disease type** by only allowing for shuffling within a given disease category, and counting the number of fusion-dependency pairings occurring by chance across all fusions with each iteration. This allowed us to build empirical null distributions for 1) the count of partner fusion-dependency pairings and 2) the count of collateral fusion-dependency pairings that would be

expected by chance across all fusions. P-values were calculated based on the number of instances in 1,000 permutations that were greater than or equal to observed counts of fusion-dependency pairings.

Cell line permutation-based FDR estimation as an approach to fusion-associated dependency discovery. Because of the non-Gaussian distribution of dependency scores and small numbers of cell lines with any given fusion, we carried out additional cell line permutation-based FDR estimation as an approach to fusion-associated dependency discovery. For each fusion, we limited hypotheses tested to a gene set of partner and other TAD genes. We then carried out 1,000 cell line permutations ***p*** based on the number of cell lines ***n*** containing a fusion in our dataset (ranging from 1-11), calculating t-statistics across all genes ***g*** for each permuted iteration. This resulted in a ***p*** **x** ***g*** permutation matrix for each ***n***. The permutation matrix corresponding to the ***n*** for a given fusion was used to calculate FDR values for the gene set as previously described(1,2). For each fusion-gene pairing, we defined **S**, the number of actual significant features, as the number of genes in the gene set with a t-statistic greater than or equal to the gene in consideration. We defined **F**, the average number of false positives, as the average number of genes in the gene set across all permutations ***p*** for which t-statistics were greater than or equal to the gene in consideration. This allowed us to calculate an FDR for each gene in the gene set for a fusion as **FDR = F/S**.

Fusion representation in clinical samples. Fusion calls from RNAseq data were available from a prior study utilizing different methodology for fusion detection among 9,624 TCGA tumor samples from 33 cancer types as previously described(3). For fusions associated with dependencies in our analysis that were derived from tumors represented in the TCGA, we evaluated the frequency that an exact fusion match was seen in the clinical dataset. For this same fusion set, we also determined which partner was the most recurrent in the clinical dataset, and the frequency with which this partner was seen (with 99% of CCLE fusions with associated dependencies having at least one partner seen in the TCGA fusion dataset).

Determining sgRNA location and illustrating fusion transcripts. CRISPR-Cas9 guide (sgRNA) location was evaluated relative to fusion breakpoints for COSMIC fusions and other fusions with associated partner dependencies. This was based on the logic that if the 5' fusion partner had forward orientation then the region to the left of the breakpoint was preserved, and if it had reverse orientation then the region to the right of the breakpoint was preserved in the resulting fusion; the opposite logic was applied to the 3' fusion partner. Based on this, we were able to determine if the 3-5 sgRNAs for a given gene involved in a fusion mapped onto the resulting transcript. Resulting fusion transcripts and relative location of sgRNAs were visualized using the St. Jude ProteinPaint web application (https://proteinpaint.stjude.org/)(4). Approximately 20% of partner dependencies were associated with fusions for which sgRNAs did not fall on the predicted transcript.

Code Availability. All analysis was done in R (v.3.6.1) using the RStudio GUI (v.1.2.5001). Select logistic regression analyses were carried out on a virtual machine with R software through the Terra Google Cloud Platform. Commercially available Adobe Illustrator 23.1.1 (2019) was used for figure formatting. Parts of conceptual diagrams were made with BioRender at https://app.biorender.com/. All of the scripts for analysis and other figure production were built in-house and are provided on GitHub at https://github.com/riazgillani/Gene-fusions-create-partner-and-collateral-dependencies-that-are-essential-to-cancer-cell-survival.

References

1. Millstein J, Volfson D. Computationally efficient permutation-based confidence interval estimation for tail-area FDR. Front Genet. 2013;4:1–11.
2. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100:9440–5.
3. Gao Q, Liang WW, Foltz SM, Mutharasu G, Jayasinghe RG, Cao S, et al. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. Cell Rep. 2018;23:227–38.

4. Zhou X, Edmonson MN, Wilkinson MR, Patel A, Wu G, Liu Y, et al. Exploring genomic alteration in pediatric cancer using ProteinPaint. Nat Genet. 2016;48:4–6.